Genomic variation and evolution of

Clostridium difficile

Miao He

Darwin College, University of Cambridge

This dissertation is submitted for the degree of Doctor of Philosophy

September 2011

Declaration

This dissertation describes my work undertaken at the Wellcome Trust Sanger Institute between May 2008 and September 2011, under the supervision of Profs. Gordon Dougan and Julian Parkhill in fulfilment of the requirements for the degree of Doctor of Philosophy, at Darwin College, University of Cambridge. This dissertation is identical to that which was examined, except as required by the examiners by way of correction.

This thesis is the result of my own work and includes nothing which is the outcome of work done in collaboration, except where indicated in the text. The work described here has not been submitted for a degree, diploma or any other qualification at any other university or institution. I confirm that this dissertation does not exceed the page limit specified by the Biology Degree Committee.

Miao He

Cambridge, September 2011

Abstract

Genomic variation and evolution of *Clostridium difficile* **Miao He**

Clostridium difficile has rapidly emerged, in part through the transcontinental spread of various PCR ribotypes including 001, 017, 027 and 078, as the leading cause of antibiotic-associated diarrheal disease in humans. In particular, a previously rare ribotype 027 was recognized as the underlying cause of a number of hospital outbreaks worldwide. However, the genetic basis of the emergence of *C. difficile* as a human pathogen is unclear.

In this thesis, comparative genomic analysis was used to identify genetic variation within the *C. difficile* population and to further understand the evolution of this organism. Genome comparison between isolates belonging to different ribotypes revealed *C. difficile* is a genetically diverse species, which is estimated to have evolved within the last 1.1–85 million years. Disease-causing isolates have arisen from multiple lineages, suggesting that virulence can evolve independently. Horizontal gene transfer and large-scale recombination of core genes have shaped the *C. difficile* genome over both short and long time scales.

Ribotype 027 isolates have a highly similar genomic backbone. To understand the genetic characteristics driving the emergence of this group and identify the genetic relationships between pre-epidemic and recent 027s, whole genome sequencing was applied to a global collection of 339 isolates spanning 25 years. Phylogenetic analysis based on SNPs identified within the core genome discriminated between >100 distinct genotypes and identified two distinct epidemic lineages that have acquired fluoroquinolone resistance independently. One of these lineages has spread more widely and contains descendants from Canada, the USA, the UK, and Australia. Further antibiotic resistance mutations and potential signatures of immune selection were also identified. Strikingly, even among these isolates, which share a highly similar core genome, there is evidence that large-scale homologous recombination and horizontal gene transfer are significant.

The global collection also included >100 ribotype 027 isolates sampled from the same English hospital and the associated patient capture areas. Phylogenetic analysis was used to distinguish relapse from re-infection cases within this particular sample set. Additionally, by combining temporal and spatial data, the use of genetic variation analysis to study local hospital transmission was explored.

Publications

Publications associated with the work described in this thesis:

He, M., Miyajima F., *et al*. Two independent fluoroquinolone-resistant lineages of epidemic *Clostridium difficile* 027/BI/NAP1 emerged in North America and spread globally. Submitted.

Castillo-Ramirez, S., Harris, S.R., Holden, M.T., **He, M.**, Parkhill, J., Bentley, S.D., and Feil, E.J. (2011). The impact of recombination on dN/dS within recently emerged bacterial clones. PLoS Pathog 7, e1002129.

He, M., Sebaihia, M., Lawley, T.D., Stabler, R.A., Dawson, L.F., Martin, M.J., Holt, K.E., Seth-Smith, H.M., Quail, M.A., Rance, R., *et al.* (2010). Evolutionary dynamics of *Clostridium difficile* over short and long time scales. Proc Natl Acad Sci U S A 107, 7527-7532.

Stabler, R.A., Valiente, E., Dawson, L.F., **He, M.**, Parkhill, J., and Wren, B.W. (2010). In-depth genetic analysis of *Clostridium difficile* PCR-ribotype 027 strains reveals high genome fluidity including point mutations and inversions. Gut Microbes 1, 269-276.

Stabler, R.A., **He, M**., Dawson, L., Martin, M., Valiente, E., Corton, C., Lawley, T.D., Sebaihia, M., Quail, M.A., Rose, G., *et al.* (2009). Comparative genome and phenotypic analysis of *Clostridium difficile* 027 strains provides insight into the evolution of a hypervirulent bacterium. Genome Biol 10, R102.

Croucher NJ, Fookes MC, Perkins TT, Turner DJ, Marguerat SB, Keane T, Quail MA, **He M**, Assefa S, Bähler J, Kingsley RA, Parkhill J, Bentley SD, Dougan G, Thomson NR. (2009) A simple method for directional transcriptome sequencing using Illumina technology. Nucleic Acids Res. 37(22):e148.

Perkins TT, Kingsley RA, Fookes MC, Gardner PP, James KD, Yu L, Assefa SA, **He M**, Croucher NJ, Pickard DJ, Maskell DJ, Parkhill J, Choudhary J, Thomson NR, Dougan G. (2009) A strand-specific RNA-Seq analysis of the transcriptome of the typhoid bacillus *Salmonella typhi*. PLoS Genet. 5(7):e1000569.

Acknowledgements

I am deeply indebted to my supervisors, Profs. Gordon Dougan and Julian Parkhill for allowing me to pursue this project, and for their guidance, advice, encouragement, patience and constructive criticism throughout my PhD. I am most grateful to my thesis committee advisors Drs. Alex Bateman, Chris Tyler-Smith, and Prof. James Wood for ideas and discussions. My gratitude also goes to the Wellcome Trust for providing for my research and living expenses.

I am very thankful to Dr. Trevor Lawley for numerous discussions on *C. difficile*, and for generously sharing his time and expertise. The mouse colonization experiment was conducted by Drs. Trevor Lawley and Simon Clare. It is also through Dr. Lawley's effort in obtaining a comprehensive strain collection that I was able to carry out the study of global BI/NAP1/027 microevolution.

I am hugely indebted to WTSI sequencing and finishing teams. Michael Quail, Richard Rance, David Harris, Elizabeth Gibson, Craig and Nicola Corton, Hilary Browne, Graham Rose, Karen Brooks, Christine Burrows, Louise Clark, Vicky Murray, Scott Thurston, Andries van Tonder, and Danielle Walker have all participated in generating the data used for this analysis. I thank all of them very much.

So many people in Teams 15 and 81 at WTSI have provided invaluable help and support to this project. The genomic DNA used for sequencing was prepared by Derek Pickard, Louise Ellison and Claire Raisen. I am very grateful to Kathryn Holt, who kindly helped me with the SNP analyses and provided ideas and suggestions when I first started. I am also thankful to Mohammed Sebaihia and Stephen Bentley for teaching me the basics of Artemis and comparative genomic analyses, and to Helena Seth-Smith, who helped to validate strain identities. A very big thank you to Simon Harris and Thomas Connor, who had generously shared their time and expertise in phylogenetics and population genetics, and to the Pathogen informatics team, including Thomas Dan Otto, Sammy Assefa and Jason Tsai, who have contributed analysis tools for sequencing data.

This project would not have been possible without the help of many researchers around the world who generously provided strains to be sequenced. Their names and affiliations are listed in Appendix A. Many thanks to Fabio Miyajima and Paul Roberts at the Liverpool University Hospital, and Brendan Wren, Richard Stabler, Lisa Dawson and Melissa Martin at London School of Hygiene & Tropical Medicine, who not only provided isolates but also shared valuable insights and information important to the project.

I also want to thank Sophie Palmer, Annabel Smith and Christina Hedberg-Delouka for looking after me while I'm writing and keeping me in check. I wouldn't be here without the support of my family and friends, whose words and presence keep me positive and sane.

To my parents

“The secret… I guess you've just got to find something you love to do and then… do it for the rest of your life.”

Glossary

ACT	Artemis comparison tool
CDI	*C. difficile* infection
CDS	Coding sequence
CGH	Comparative genomic hybridization
contig	Contiguous sequence from overlapping reads
GA	Billion years ago
GTR	General time reversible substitution model
homosplasy	Similarity through convergent evolution but not by descent
ICE	Integrative and conjugative elements
IS	Insertion sequence
JC	Jukes and Cantor substitution model
MLST	Multi-locus sequence typing
MLVA	Multiple-locus variable-number tandem-repeat analysis
MRCA	Most recent common ancestor
PFGE	Pulsed-field gel electrophoresis
PMC	psedomembraneous colitis
REA	Restriction endonuclease analysis
RFLP	Restriction fragment length polymorphism
tMRCA	Time to most recent common ancestor
WTSI	Wellcome Trust Sanger Institute

List of Figures

1.1 Phylogenetic tree of *C. difficile* and other *Clostridium* species. **.....3**

1.2 Distribution of BI/NAP1/027 *C. difficile* in European countries as of June 2008. .....**7**

1.3 Distribution of BI/NAP1/027 *C. difficile* in the USA and Canada. **8**

1.4 *C. difficile* pathogenicity locus and domain organization of toxin B. **15**

1.5 Mechanisms for DNA transfer in bacteria. **25**

1.6 Models for bacterial population structure shaped by selection and demographic processes. **29**

1.7 Number of sequenced complete genomes in each year from 1995 to 2010 (Genbank). **39**

1.8 Increase in sequencing capacity during the first decade in 21st century. **40**

1.9 Mapping paired-end reads to a reference sequence and identify SNPs... .....**45**

2.1 Phylogenetic trees of *C. difficile* based on whole-genome sequences. **60**

2.2 Phylogenetic tree of a diverse collection of *C. difficile* isolates based on concatenated non-recombining core CDSs. **61**

2.3 SNPs between CD196 and 24 other hypervirulent *C. difficile* isolates. **64**

2.4 Bayesian skyline plot (group number = 10) shows a recent population expansion of the hypervirulent group. **65**

2.5 Genomic islands within isolates of the hypervirulent clade. **66**

2.6 Comparison between parts of the *C. difficile* M120 genome and the genome of *S. pyogenes* MGAS10750. .....**68**

2.7 Comparison between parts of the *C. difficile* M120 genome and the genome of *Thermoanaerobacter* sp.X514. ....**69**

2.8 Signature of recombination in the deep-branching phylogeny. **71**

2.9 Trajectory of 1/(dN/dS) within the *C. difficile* phylogeny over time. **73**

2.10 Changes in core- and pan-genome size of *C. difficile* in relation to the number of genomes. **76**

2.11 Distribution of orthologues CDSs in *C. difficile* strains 630, CD196 and R20291. ....**77**

2.12 Circular representations of *C. difficile* chromosomes. **79**

2.13 A conjugative transposon carried by R20291 but absent from CD196.**81**

3.1 Methods for assessing accuracy in short reads mapping and SNP detection. **90**

3.2 Numbers of false positive SNPs and percentage of false negative SNPs in relation to sequencing data coverage and different SNP filtering and validation measures. **94**

3.3 Global phylogeny of *C. difficile* BI/NAP1/027. ....**97**

3.4. NeighbourNet network of the *C. difficile* BI/NAP1/027 lineage. **98**

3.5 Expansion of clade 1 from the global phylogeny. **99**

3.6 Expansion of clade 2 from the global phylogeny. **101**

3.7 Expansion of clade 3 from the global phylogeny. **102**

3.8 Expansion of clade 4 from the global phylogeny. **103**

3.9 A summarized tree based on sampled trees from a BEAST run with a relaxed log exponential clock model. **105**

3.10 SNPs between R20291 and eight *C. difficile* isolates showing homologous recombination blocks. **109**

3.11 Isolates and lineages carrying substitutions conferring resistance to antimicrobial drugs. **112**

3.12 Bayesian skyline plot indicating changes in population size of this BI/NAP1/027 sample collection. **115**

3.13 Genomic regions that are present in the reference (M7404) but absent from isolates in clade 1. ..... **116**

3.14 Comparison of three versions of CTn*CD5* in strains R20291, M7404, and 2007855. ..... **117**

3.15 Genes gained along the branch connecting clades 1 and 2, shown in the genome of M7404. ..... **118**

3.16 Accessory genome components of isolates in clade 1. ..... **120**

3.17 Accessory genome components of isolates in clade 2. ..... **121**

3.18 Accessory genome components of isolates in clade 3. ..... **122**

3.19 Accessory genome components of isolates in clade 4. ..... **123**

3.20 Comparison between phi*CD38-2* and unique contigs from isolate BI-2a. **124**

3.21 Comparison between a contig in Gla007 and CTn*4* in 630. ..... **125**

3.22 Comparison between a contig unique to Cam036 and CTn*3* in 630. **125**

4.1 The pattern of long-term infection of C3H/HeN mice with *C. difficile* strain BI-7 (ribotype 027). **136**

4.2 Un-rooted maximum likelihood phylogeny of Liverpool isolates. **139**

4.3 Temporal graph of *C. difficile* ribotype 027 genotype assignment from the same patient at multiple infection episodes. **141**

List of Tables

1.1 Scientific classification of *Clostridium difficile.* **2**

1.2 Comparison of sequencing platforms. **41**

2.1 Sequence coverage of the broad collection of *C. difficile* isolates. .....

..... **58**

2.2 Details of the hypervirulent isolates included in this chapter. **62**

2.3 Potentially positively selected genes in *C. difficile.* .....

..... **75**

3.1 Discriminatory SNPs and predicted functional effects within the

BI/NAP1/027 *C. difficile* lineage. **107**

3.2. Homoplasic SNPs and associated gene products. **111**

3.3. Estimated mutation rates using a maximum likelihood expansion model.

..... **113**

3.4 Fourteen coding sequences gained along the branch connecting clades 1

and 2. ..... **119**

4.1 Number of *C. difficile* isolates per semester by ribotypes. **137**

Contents

Abstract iii

Publications iv

Acknowledgements v

Glossary viii

List of Figures ix

List of Tables xii

1. Introduction 1

1.1 *C. difficile* 1

1.1.1 The bacterial species *C. difficile* 1

1.1.1.1 Classification 2

1.1.1.2 History of *C. difficile* discovery and research 3

1.1.2 *C. difficile* infection 4

1.1.2.1 Symptoms 4

1.1.2.2 Diagnostics 5

1.1.2.3 Epidemiology 6

1.1.2.4 Community-associated CDI 9

1.1.2.5 Environmental and zoonotic *C. difficile* 9

1.1.2.6 Treatment, antibiotic use and resistance 10

1.1.2.7 Typing schemes for *C. difficile* 13

1.1.3 Prominent virulence factors and transmission agent 14

1.1.3.1 *C. difficile* toxins 14

1.1.3.2 *C. difficile* spores, spore formation and germination 16

1.1.4 *C. difficile* genomics and genetic diversity 18

1.2 Genetic variation and evolution of bacterial populations 19

1.2.1 Genetic diversity and evolution of bacterial populations 20

1.2.2 Mechanisms that generate genetic diversity in bacteria 21
1.2.2.1 Nucleotide substitution 21
1.2.2.2 Horizontal gene transfer 23
1.2.2.3 Recombination 28
1.2.3 Impact on speciation 32
1.2.4 Considerations in studying bacterial populations 34
1.2.4.1 Typing schemes and choice of genetic loci 34
1.2.4.2 Sampling of bacterial pathogens 36
1.3 Genome sequencing of bacterial pathogens 37
1.3.1 Next generation sequencing 39
1.3.1.1 The new sequencing technologies 40
1.3.1.2 Next generation sequencing bioinformatics 43
1.3.2 Studying bacterial populations using next-generation technologies 45
1.4 Thesis outline 47
2. Genomic variation of *C. difficile* over short and long time scales 49
2.1 Introduction 49
2.2 Materials and methods 51
2.2.1 Bacterial isolates 51
2.2.2 DNA sequencing and assembly 52
2.2.3 Genome annotation, comparison, identification of orthologues and unique genomic regions 53
2.2.4 Phylogenetic analyses 53
2.2.5 SNPs detection and CDS alignments 54

2.2.6 Recombination and selection analysis 55

2.2.7 Estimates of age and population size 55

2.2.8 Identifying non-recombining core coding sequence 56

2.3 Results 57

2.3.1 Macroevolution of the *C. difficile* species 57

2.3.1.1 The deep-branching phylogeny.....57

2.3.1.2 The age of the *C. difficile* species 61

2.3.2 Microevolution within the hypervirulent clade.....62

2.3.3 Extensive role of horizontal gene transfer in *C. difficile* evolution 65

2.3.4 Large chromosomal regions exchanged by homologous recombination.....69

2.3.5 Selective forces acting upon the *C. difficile* genome.....72

2.3.6 Core-genome and pan-genome sizes of *C. difficile* 75

2.3.7 Genome comparisons between isolates CD630, CD196 and R20291.....77

2.4 Discussion.....82

2.4.1 Insights from deep-branching phylogeny 82

2.4.2 The relative impact of recombination versus mutation.....83

2.4.3 Genomic island of potential functional impact.....83

2.4.4 Core- and pan-genome sizes.....84

3. Microevolution and global transmission of *C. difficile* BI/NAP1/027 85

3.1 Introduction 85

3.2 Materials and Methods.....87

3.2.1 Bacterial isolates 87

3.2.2 Sequencing, mapping, and SNP detection 88

3.2.3 Assessing accuracy of SNP detection 89

3.2.4 Phylogenetic analysis.....91

3.2.5 Core and accessory genome 91

3.2.6 Identification of homoplasic characters and homologous recombination.....92

3.2.7 Population history and mutation rate.....92

3.3 Results 93

3.3.1 Assessment of SNP detection method.....93

3.3.2 Phylogenetic relationship 95

3.3.2.1 Maximum-likelihood phylogeny and phylogenetic networks 95

3.3.2.2 Phylogenetic relationships inferred with Bayesian analysis 103

3.3.3 Discriminatory SNPs and potential functional consequences105

3.3.4 Signatures of recombination 108

3.3.5 Homoplasic SNPs and convergent evolution 109

3.3.6 Estimating mutation rate 113

3.3.7 Population history inference.....114

3.3.8 Horizontal gene transfer in BI/NAP1/027 lineage.....115

3.3.9 PaLoc region variation 126

3.4 Discussion.....126

3.4.1 Two lineages associated with fluoroquinolone resistance 126

3.4.2 Other insights drawn from the phylogeny.....127

3.4.3 Agreement with earlier analyses based on a smaller sample set 128

3.4.4 Antibiotic resistance 128

3.4.5 Genetic changes underlying the success of BI/NAP1/027 129

3.4.6 Potential mechanisms for large chromosomal region replacement 130

4. Hospital transmission and persistence of *C. difficile* from a whole genome sequencing perspective 132

4.1 Introduction 132

4.2 Materials and methods.....134

4.2.1 Bacterial isolates 134

4.2.1.1 Ribotype 027 isolates from patients 134

4.2.1.2 *C. difficile* BI-7 (ribotype 027) isolates obtained over time during a mouse colonization experiment.....135

4.2.2 DNA preparation, sequencing, reads mapping and SNP detection 136

4.2.3 Phylogenetic analysis.....137

4.3 Results 137

4.3.1 Genetic diversity and microevolution of hospital ribotype 027137

4.3.2 Colonization by *C. difficile* 027 in patients.....140

4.3.3 Spatial and temporal distribution of ribotype 027 genotypes 142

4.3.4 Genetic diversity of ribotype 027 during colonization of mice145

4.4 Discussion.....146

4.4.1 Strengths and limitations of spatial temporal genotype analysis 146

4.4.2 Relapse, re-infection and multiple strain carriage 147

4.4.3 Insights from *C. difficile* colonization in mice.....148

5. Final discussion 149

5.1 Significance of homologous recombination 149

5.2 Insights from the study of a global collection of BI/NAP1/027 150

5.3 Selective pressure and gene candidates for functional study.....151

5.4 Insights from local hospital transmission study.....151

5.5 SNPs as genetic markers for genotyping.....152

References 153

A. BI/NAP1/027 global collection used in Chapter 3 181

B. BI/NAP1/027 local hospital collection used in Chapter 4 193